Open Research Pilot Project Meeting 13th September 2017

Held at the Wellcome Trust, London

What Was Discussed

1. Welcomes

Lauren Cadwallader welcomed the group. Round-the-table introductions followed. The four University of Cambridge Open Research Pilot Project groups were represented at the meeting (Dr Marta Costa, Dr Laurent Gatto, Dr Greg Jefferis, Dr David Savage and Dr Ben Steventon). Also present were Robert Kiley and David Carr from the Wellcome Trust (WT) Open Research team, as well as Dr Danny Kingsley, Dr Lauren Cadwallader, Dr Marta Busse and Dr Debbie Hansen from the University of Cambridge Office of Scholarly Communication (OSC) and Georgina Cronin from The Betty & Gordon Moore Library.

2. Activity – Working Openly and the Research Lifecycle

Attendees split into small groups and considered each of the six research workflow phases defined in one model of the research lifecycle (the Utrecht diagram): Discovery, Analysis, Writing, Publication, Outreach, Assessment. Firstly, the groups defined actions: how they are currently or could be working openly in each of the phases. Secondly they considered the tools required to do the actions. The actions from all groups were placed on large sheets, grouped together by similar actions, and the tools were attached to relevant activities. A brief discussion identified which activities were possible, and what some challenges were for those that were not. The results are in the Appendix.

3. Presentations from Each of the ORPP Groups

3.1. Dr Marta Costa and Dr Greg Jefferis (Dr Lauren Cadwallader)

3.1.1 Presentation

Greg and Marta briefly described two projects relevant to the Open Research Pilot Project.

Virtual Fly Brain is an interactive on-line tool to make it easier for researchers to find relevant information about the fly brain. The benefits of a tool such as this include image and data sharing, standardisation in the way that people talk about things pertaining to the fly brain, ready access to data and literature for international collaboration as well as private research. The challenges associated with this tool include sizes of files (terabytes associated with image data) and future funding of this resource.

For the Connectomics project (reconstruction of the circuits involved in olfactory learning and memory in the adult fly), Marta reported that, in looking at ways to share information between the group, they took part in an Electronic Lab Notebook (ELN) trial. They looked at 2 ELNs, but found that neither fitted with their current workflow.

3.1.2 Group Discussion

The main discussion centred around long-term data preservation (how long? 20 years?), FAIR (Findable, Accessible, Interoperable, Reusable) principles around the data, and the balance between making data accessible to as wide an audience as possible whilst still retaining usability for the specialist (i.e. How open? Open for who?).

3.1.3 Issue for Later Discussion: Sustainability – funding and the digital resources.

3.2. Dr Ben Steventon (Dr Marta Busse)

3.2.1 Presentation

Ben described imaging of zebrafish. Bespoke microscopes are required for imaging different parts of a zebrafish embryo. The movies jerk and the part of interest moves out of view. There is new software to analyse tracking data. There is a need to share the training required to use the analysis software and data. The associated imaging and tracking files are very large (e.g. 3D movies ~ 2 terabytes; compressed data ~80-100 gigabytes; tracking data and projected movie ~ 1 gigabyte. One question is how the data can be made available to others (e.g. for others to continue investigations).

3.2.2 Group discussion

The main discussion centred around the data: e.g. the storage of the raw data (size and sustainability), where the data should be made openly available, availability of specialized software for re-use of the data and sharing the training for using the data (as this is complicated). The preferred solution is a dedicated platform with the software allowing for re-use of images for various projects (Image Data Resource, https://idr.openmicroscopy.org/about/). This will be investigated further with the platform developers. Other options put forward by the group were the Cambridge repository (but currently problems with versioning), Figshare, the Jisc repository of choice). Another suggestion was to explore collaboration with companies that are used to dealing with big data – perhaps they could provide solutions or grants for investigation or other resources (as a caveat it was thought that whether the existing funder would need to agree to this would need to be explored).

3.2.3 Issue for Later Discussion: Accessibility of data

3.3. Dr David Savage (Georgina Cronin)

3.3.1 Presentation

Research in the area of human genetics has been conducted in an open way for many years. Biobank has been open for a long time. The benefits have been huge. For example, studies can be done on much larger populations than previously practical due to open data. The main concern here is scientific ethos. In this field, industry-lead research is rigorous as each step depends on the outputs of the last. In the academic environment, the main outcomes are publications. In this environment there are very intelligent competitive researchers and there is huge competition for funding. In the quest

for prestige, there is concern that scientific rigour may be lost. If results are not validated, this could be wasteful. In journals such as eLife, where there is a move away from conventional peer review, there may be problems with bias creeping in. Older researchers have long-established colleagues in the field who may review for each other. Also, would younger researchers want to be seen as potentially antagonising long-established and highly regarded personalities in their field? What would be the right career motivation instead?

3.3.2 Issue for Later Discussion: Reproducibility and integrity, incentivising researchers.

3.4. Dr Laurent Gatto (Georgina Cronin)

3.4.1 Presentation

In preparation for this meeting, Laurent went back to two earlier blog posts: Open Research Project, first thoughts and Exploring and sharing spatial proteomics data.

Laurent explained about the tool Spatialmap.org (the subject of the second blog post). This is the first implementation of a prototype software tool, developed by a student, which enables proteomics data to be shared through graphical interfaces. This tool can be used to search data, link with other resources and share data with a reviewer, for example. In order to further develop this tool, funding would be needed, but in order to get funding, would the tool need to be made more general?

In the first blog post, the following questions were asked: How can Open Research be promoted? What are the incentives? What are the rewards (do we need threats?)? An individual can invest their own time and effort into trying to do the right thing, but who are the right people to drive the change? Tools are not the bottle neck. The issue is how we can promote and encourage open science to be done by everybody. Early career researchers should be given incentives to work openly.

3.4.2 Group Discussion: One question that was raised was are institutions also to blame for the inertia? It was commented that institutions are run by managers and senior scientists who have succeeded through a particular way of working. It can become frustrating and demotivating to try to make changes and not succeed. Further, the comment was made that a change can be stopped if one senior person, for example, doesn't understand or believe in the need for the change. Early career researchers are very active, but is it their job? They have to fight particularly hard as it is to survive in this environment.

3.4.2 Issue for Later Discussion: How can we promote and encourage open science?

3.5 Wellcome Trust

Robert Kiley and David Carr gave a progress report, and talked about the WT open publishing platform, Wellcome Open Research, and other ways open science is being progressed by Wellcome Trust.

They talked about their initial ambitions for this Open Research Pilot Project. They saw this as a way of promoting research outputs to be shared and a way of gaining a better understanding of the

barriers to working openly. They also want to look at how the incentives can be addressed. It was commented that no group taking part in the pilot project have submitted anything to the publishing platform, Wellcome Open Research, and the group were asked if there was a reason for this and whether anyone had any concerns. No concerns were raised by the group and it was noted that Laurent has written a peer review for a submission to this platform.

Robert commented that the Open Research Pilot Project blogs have been useful and the Wellcome Trust Open Research team agree on the key issues raised (e.g. issues associated with large files, sustainability and how to motivate on open science practices).

Robert gave a brief report about Wellcome Open Research. In less than a year, more than 100 articles have been published on this platform. At the start of the project, the project team defined what they thought success would look like. Success was defined as 25-30 publications in a year. They are therefore very pleased by the result. A range of outputs have been published. Half are research articles, and the rest are other output types such as data notes, software studies and protocols. Also, submissions have come from a broad range of institutions. They looked at how popular the platform is by comparing the volume of Wellcome Trust funded publications across the range of publications. Wellcome Open Research was found to be the 4th most used after Scientific Reports, PLoS ONE and Nature Communications. The average cost of publication on the Wellcome platform is of the order of £700, which is significantly lower than the average APC cost (of the order of £2000). Other benefits of using this platform were illustrated by an example whereby in a 3-week period, an item was submitted, reviewed, approved, published and made discoverable and then requests were received for the data. Gateways can be created where all content can be put together – the only criteria being that at least 5 articles must be attached on day 1. Gateways provide personalized portals for institutions or organizations, with links to other resources. There has been competition by some overseas centres to get a Gateway!

David reported that in July Wellcome Trust published a new policy on output management. This will clarify their expectations with regard to what should be shared and when. For example, data and software underlying publications should be made available at the point of publication. They are keen to develop guidance underlining the policy. Danny suggested that the OSC organise an "In conversation with the Wellcome Trust" event for University of Cambridge Wellcome Trust funded researchers for informing and educating about this new policy.

David and Robert see Wellcome Open Research as a clear focal point in encouraging the sharing of research outputs. They are working on a plan for the coming year. They want to look at other mechanisms and ideas for open research and sustainability.

4. Discussion

The discussion was planned around two main categories: Funding and Sustainability, and Open Research and Culture. The main points discussed are captured in Sections 4.1. and 4.2. Other business raised is in Section 4.3.

4.1 Open data, Funding and Sustainability

MRC Laboratory of Molecular Biology (LMB) sponsored one data resource for a while (someone was paid to manage the resource) but this came out of an individual researcher's project as it was considered to be of interest. Otherwise MRC does not fund data storage of any kind in the LMB.

An example of a funding and sustainability problem was described (see: http://www.biorxiv.org/content/early/2017/04/27/110825). The Human Frontier Science Program (HFSP) is an international body which awards funds for frontier research in the life sciences and is based in Strasbourg. A problem they see at the moment is that the community is reliant on core data resources. These tend to be funded by a small number of funders but the data are key to the whole community. When funders are asked for further funding to continue the resource, there is competition with new requests to fund new ideas. Also, is it the funder's job to provide long-term sustainability? The next task of the group will be to come up with a business plan as to how to take this forward. It is encouraging that this issue is being discussed at the international level.

The point was raised that there is a mismatch between a reliance on certain resources on the one hand, but a reluctance to fund for long-term sustainability. For example, when arXiv (an e-print service, operated by Cornell University, in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics) asked the physics community to provide support, they thought that it was the library's responsibility to provide funding, not the physics community. Similarly, Canadian researchers rely heavily on GenBank, but no Canadian funding agencies contribute to the costs of this resource.

There was a study (by Elixir) to define criteria for so-called core data resources (see https://f1000research.com/articles/5-2422/v2). An indicator could not be number of hits alone, for example, because disciplinary differences would need to be taken into account. One concern that was raised was that only those resources that are more community-oriented would be supported into the future; how to deal with smaller communities where data is important but to a smaller number of researchers was seen to be an issue.

It was suggested that there would need to be a mechanism that evaluates how resources are doing over time and also a need for a project 'end of life' mechanism.

Re-use of data is challenging yet beneficial. How can people be helped to use existing resources and maximise the benefits? WT groups could meet other WT groups and help spread the word. Is there a percentage that funders think ought to be spent on helping people use existing data? What other mechanisms could be used to demonstrate to the benefits of working openly?

One question posed to the researchers in the group was: what support from their institutions and funders would they want to make their data more accessible? It was commented that time was a big issue. For example, it takes time to annotate data sets to be made useable to others. One group said that they could write protocols and a series of articles to put on Wellcome Open Research, which would be a good thing, but it would take the team a long time. There is seen to be a need for dedicated staff to manage sharing of research at the institute or department level. It was commented that having had access to library colleagues through this Open Research Pilot Project has been useful for figuring out where to put their research data – could this be continued? An action was taken for the library component of the group to think about what support is being provided in this context (and into the future).

4.2 Open Research and Culture

The points raised in the discussion have been summarised under the following categories: the current system; rewards and incentives; leadership; integrity and reproducibility; funders, publishers and open access; and an open research future

The Current System

In the current climate, it was pointed out that it is easy to identify a successful researcher. A successful researcher has prizes, publishes in particular journals with high impact factors and has grants and funding. However, it was noted that nowadays a successful researcher and a good researcher are not necessarily the same thing.

Rewards and Incentives

For publications, one block to Open Access has been powerful and rich publishers. For Open Research, it is thought that the block will not be from publishers, but from the research community itself. For example, the current incentive and reward structure is a barrier to change and there is a need to re-define what is valuable and this may need to be defined at the discipline level.

One suggestion was to look at applying caveats to grants or Fellowships. For example, funders could only provide Fellowships to institutions or departments that guarantee they will provide tenure track to the receiving person at the end of 5 or 8 years. Or, only provide grants or Fellowships to institutions or departments that have signed or support the Declaration on Research Assessment (DORA). A discussion then ensued over the fact that signing and implementing were two different things and whether it would be possible to verify whether what had been agreed had been implemented. With Open Access publications, verification can be done. WT are now starting to look at how to assess end-of-grant reports. This might involve, for example, looking at the original data sharing plan and what was subsequently done about it.

WT described an initiative they will be trialling. This is a data re-use prize. Experience has shown that data which is made available often is not re-used to the extent expected (for example, the Clinicalstudydatarequest.com platform receives only 100 access requests per annum, and only 2-3 papers have resulted from data re-use. In order to encourage data re-use, WT will offer a prize to reward the most innovative and impactful re-use of a defined data resources. One action suggested was to look at whether the Australian National Data Service have done similar work.

Another suggestion was to have travel fellowships so the recipient could travel to another laboratory, where similar research was in progress, and spend a week there, for example, to explain to others how to use the data.

WT took an action to explore the idea of travel fellowships awards for good Open Research practices. WT will launch an enhanced funding mechanism for Open Research later this year. WT researchers will be able to apply for supplementary awards for open research activities. This could support travel for this purpose, as well as, for example, the use of data visualisation tools. One point raised regarding this was that credit should be given to the individual winning the award and not the head of the

laboratory who was the recipient of the original grant (as, from a process perspective, the funding would be an add-on to the grant). The recipient should be named and given the money so that they can add the achievement to their CV.

Communication was felt to be important. For example, to communicate to the community about what happens if policies are not followed. Provide incentives and rewards but also consequences if people don't comply. By 'fighting on different fronts', slowly the research environment will be able to change.

Leadership

It was felt that driving the change to Open Research requires strong leadership, and the leadership would need to come from funders and institutions for the researchers to align themselves with working openly. If funders communicated to institutions that they would only support open research, for example, institutions would have to be aligned and then the researchers. However, it was commented that this idea of forcing the issue was simplistic – what does 'My institution' mean as the University of Cambridge, for example, is so diverse and who would have the power to make those decisions on behalf of the whole research community. There is understanding and support among the senior leadership of the WT for Open Research. Their data sharing and open access policies are getting firmer. It was commented that institutions will not do things that would jeopardise funding, so change needs to be led by the funders!

Integrity and Reproducibility

How to incentivise scientific integrity and reproducibility of results was discussed. One option might be to cut funding if reproducibility could not be proven (e.g. by publishing protocols and have peer review of doing as well as peer review of reading). It was acknowledged that this would take more time. Some publishers are developing a reproducibility stack whereby, when a paper is published, for example in eLife, the software and data are made available too and a reader can mess around with the variables. If this were built into the publication process like this, monitoring reproducibility may be achievable.

However, it was commented that such tools have already been in place for a decade or more (maybe not in such a nicely fine-tuned way) and why should the research community need the publishers to be involved in this? One worry if publishers involve themselves in Open Research, they would do it the way they want to do it and it will cost the research community. It was suggested that publishing was done instead on their 'own' platform, although it was commented that even the 'own' platform would have a commercial backdrop. However, the benefit is if data is archived it will not disappear.

Funders, Publishers and Open Access

Funders could insist that a green open access route is followed or pay for gold open access for only truly open access journals. Currently, ~71% of the £4 million invested each year by WT in OA is with hybrid publishers – the researchers are choosing where to publish, not the funders. This is likely to be because of the impact factor, prestige and reputation of the journal. There was a feeling that WT grant panels and the people on the peer review panels, still look at journal title as a proxy for quality of an output. Could the emphasis on impact factor be stopped? Could all WT researchers be forced to publish in Wellcome Open Research?

If WT stopped paying for open access in hybrid journals, would researchers stop applying for funding? The feeling was no: they will be looking for other funding in future so would still want to publish in journals like Cell for example. Researchers perceive that when applying for grants, their record for publications in particular journals is very important. The ethos needs to change with regard to the need to publish in particular journals.

Alternatively, could the value given for APCs be limited, say to $1000? This would need consensus across all major funders internationally to implement. This would be hard. For example, CERN worked on a contract, with the American Physical Society, for making the top 12 journals used by the high energy physics community open access. All relevant institutions and associated libraries had to chip in. It took 5 years to get a deal in place.

Open access could be a vehicle for improving open research. If payment for articles in hybrid journals, for example, was stopped, the money saved could be used to invest in other aspects of open research.

One other comment was that open access is not just the freedom to read, but the freedom to mine and the freedom to publish: this possibility is currently only for rich institutions.

An Open Research Future

One question that was raised was whether the group thought that Open Research was an evolutionary strategy. For example, would we look back after 5 years and say that researching openly was clearly beneficial, or will it continually require effort and funding? Will the issues we are seeing at the moment dissolve in 5 years because all will be doing the right thing?

4.3 Other Business

WT asked if they could use this group as a sounding board for an initiative to look at whether there is a need for a WT-branded repository. They would be looking to ask the group to review the draft requirements (to provide a sanity check). One initial thought was that merging analysis and archiving could be beneficial but this would likely be domain-specific: i.e. a platform where data is loaded, worked on and then archived (versioning would be critical). The OSC could be involved from a best practice point of view (e.g. certification, data seal of approval). An action was taken (OSC) to look at whether there were findings from the Jisc shared repository pilot and whether these could be shared with the group.

Finally the comment was made that all challenges raised have already been discussed openly by other groups and some solutions put forward by them. It was suggested that when new policies are being released, discussions should be broadened to the community at large. Many people have thought about these problems.

Appendix

Activity – Working Openly and the Research Lifecycle

Themes: Discovery, Analysis, Writing, Publication, Assessment, Outreach

1. Discovery ELN lack of Institutional support

Possible Actions (Green)	Potential Tools to link to the Actions (Pink)
Ensure outputs are accessible via PubMed & Google	PubMed for data? Joined up search ability to find all research, not just what one publisher owns
ELN previous research?	Github et al, Open Portals, shared & generic resources
Previous outputs from research discoverable (journals, data, software, blog posts, internet, community outputs)	Data sharing, share ideas and project aims with colleagues, e.g. conferences
Provide excellent metadata	Fair data principles
Open journal clubs	Text Mining
Use a machine readable licence	TM
Create videos and/or interactive content about work	Online courses (training)
Use existing resources, hosting/data tools	Integration or data sharing initiatives in conferences
Website, web supplier data-methods	Public contributions to research
Develop resources to share processes or how to with interested parties	Domain specific resources, e.g. Virtual Fly Brain, Repositories

2. Analysis

- **Interim –** concern re quality of data scooped
- **Replication –** time & money lack of incentive if not good protocol
- **Logistic -** problem of sharing large data sets

Possible Actions (Green)	Potential Tools to link to the Actions (Pink)
Interim analysed data	Github et al
	Shareable software tools > central management
	OSF
	Data repositories (PRIDE)
	Git & Github, R + Rmerkdown, Jupyter (multilingual)
	Ways of caching intermediate data?
Share open source code	e-lab books, shareable lab notebooks, data annotation ELNs
Sticking to open source tools	R et al
Reproducible analysis for projects	Open protocols
Reproduce others results, replication	Open MTA? Lab website development?
Publication of other outputs- data, analysis, pipelines	Standardised metadata (data description)
Open protocols	Metadata documentation schemes
Publish code and intermediate analysis in Github, Figshare etc	Standards on metadata, standard software for data type

3. **Writing**
    - **Writing openly** relying on 3rd party XXX?.features
    - **collaborators –** not keen
    - **lack of maturity of tools**

Possible Actions (Green)	Potential Tools to link to the Actions (Pink)
ELN	Tools to connect between Word editing programs
Collaborative writing tools	Collaborative editing –Gdocs, Overleaf
Writing in the open	Tools linked to open journals e.g eLite
	Googledocs and Paperpile
	Authorea?
	Ovderleaf/workplace tools that produce machine readable outputs
Post all/full revisions of paper	Github et al
Share drafts reprints	
Utilize pre-print to share work at early stage	
Blog regularly about work	

4. **Publication**
    - **OA** career strategic problem
    - **Limits collaboration**

Possible Actions (Green)	Potential Tools to link to the Actions (Pink)
Cite your data properly	Data repositories (PRIDE, Zenido)
	Don't limit references, index SI citation
Apply an open licence to all outputs	
OA journals, OA only, pre-prints	Open Access Service@cambridge
	Open peer review platforms
	ArXiv
	Preprint servers
	Availability of journals> high quality services/reasonable price
Publish full range research outputs +ive & -ive on OA platforms	Github et al
Deposit materials in institutional repository for re-use	Research data service @cambridge
	Free yet established &/or sustainable publishing platforms without any conflict of interests (commercial etc) Elsevier?
	Shared manuscript standards across publishers-file types, metadata

5. **Assessment**
    - **Need to capture other outputs**
    - **Improve what content tools do**

Possible Actions (Green)	Potential Tools to link to the Actions (Pink)
Recognise track record in Open Research in funding/career decisions	Grant focus on science not journal IF
Theses	ORCID/Google scholar with data integration
	Guidelines for assessment which reflect DORA principles
	ORCID with new functionalities
Open Peer review, post public per review	Open peer review
Support open peer review	Post pun review biorXiv
How about open peer review of grants	University policy on assessment
	Ways to measure research evaluation/comment online
Eschew IF journal proxies	DORA, Github et al
	Open badges
	Institutional recognition through visible reward systems that are more subtle than simply promotion (badges etc)

6. **Outreach**
    - **Time hungry**
    - **Interest level from audience**
    - **Not rewarded**
    - **Where put plain English**
    - **Limited value to public**

Possible Actions (Green)	Potential Tools to link to the Actions (Pink)
Community meetings & presentations	Database of contacts, schools, museums etc
Encourage open practices with community and beyond ADVOCATE!	PE policy > mandatory
National clinical service	Weblink/blog server
P support group	Training in webpage design, project or personal webpages
Conference presentations	Data repo (OmicsID…)Github et al
Social media for sharing collaboration	Twitter, Jekyl (bloggy)
Media training	Social media tools (Twitter, Wordpress etc
Schools partnerships	
Social networks	
Blog, Tweet	
Utilise sharing platforms such as ‘slideshare’ to disseminate work to broader audiences	